Comparison of statistical models performance in case of segmentation using a small amount of training datasets

François Chung, Jérôme Schmid, Nadia Magnenat-Thalmann, Hervé Delingette

The Visual Computer manuscript No.
(will be inserted by the editor)

Comparison of statistical models performance in case of segmentation using a small amount of training datasets

François Chung[1], **Jérôme Schmid**[2], **Nadia Magnenat-Thalmann**[2] **and Hervé Delingette**[1]

Received: date / Accepted: date

Abstract Model-based image segmentation has been extensively used in medical imaging to learn both shape and appearance of anatomical structures from training datasets. The more training datasets are used, the more accurate is the segmented model as we account for more information about its variability. However, training datasets of large size with a proper sampling of the population may not always be available. In this paper, we compare the performance of statistical models in the context of lower limb bones segmentation using MR images when only a small number of datasets is available for training. For shape, both PCA-based priors and shape memory strategies are tested. For appearance, methods based on intensity profiles are tested, namely mean intensity profiles, multivariate Gaussian distributions of profiles and multimodal profiles from EM clustering. Segmentation results show that local and simple methods perform the best when a small number of datasets is available for training. Conversely, statistical methods feature the best segmentation results when the number of training datasets is increased.

Keywords model based segmentation · statistical models · principal component analysis · clustering

[1] François Chung and Hervé Delingette
Asclepios Research Team, INRIA Sophia-Antipolis
B.P. 93, 2004 Route des Lucioles
06902 Sophia Antipolis, France
E-mail: francois.chung@inria.fr
Tel.: 0033-(0)4 92-387160
Fax: 0033-(0)4 92-387669

[2] Jérôme Schmid and Nadia Magnenat-Thalmann
MIRALab, University of Geneva
Battelle, bâtiment A, 7 Route de Drize
1227 Carouge/Genève, Switzerland
E-mail: jerome.schmid@miralab.ch
Tel.: 0041-(0)22 37-97769
Fax: 0041-(0)22 37-90079

1 Introduction

Building meaningful priors for model-based image segmentation purposes is an important topic in medical imaging. To account for a large variety in both shape and appearance, numerous datasets are usually required in a training stage. Atlas-based methods are designed in the hypothesis that a significant number of training datasets may yield a reasonable mean model (*i.e.* in the sense of a population mean) as well as meaningful modes of variation. Those methods usually define a reasonable estimation of the prior using the principal modes of variation (*i.e.* variations may be controlled with few parameters). They have been successfully used in coarse-to-fine approaches where the algorithm starts with a rough approximation of the prior (*i.e.* with few degrees of freedom) and evolves adding more variation until reaching a steady state [14, 19].

In the literature, shape variations are often described using Principal Component Analysis (PCA) [19, 22, 14, 2, 4, 6] of Point Distribution Models (PDM) [5]. From a sufficient number of shapes with known point correspondence, a mean shape is calculated. During the segmentation, a mesh initialized as the mean shape is deformed until it fits a desired target shape. PCA-based methods allow the mesh to be deformed only in the range of estimated variations. In this way, deformations occur in a predictable way since they originate from the principal variations observed from the datasets used to build the statistical model. Regarding the appearance, methods such as Active Shape Models (ASM) [6] and Active Appearance Models (AAM) [7] were proposed to account for the main intensity variation around and within structures of interest respectively.

However, the number of datasets required to account for both the shape and appearance of a structure may be an issue. First, medical imaging acquisitions require time and resources, and this may limit the number of datasets available

for training. Second, the large number of acquisition protocols and hardware characteristics (especially in case of versatile modalities such as MRI) tend to produce images with a large variety of intensity distribution for the same structure of interest. Additionally, noise and artifacts (*e.g.* patient movement and partial volume effect) are likely to corrupt the image intensity. Those factors strongly affect the construction of priors, as they bring meaningless intensity information into the appearance priors. Also, the manual segmentation of images by an expert, which is required for the extraction of shapes, is a tedious task and is a limiting factor for the availability of training datasets. Finally, the large natural variability of shape and appearance cannot be well represented by a Gaussian distribution assumed by PCA and thus capturing all the shape variations is still very challenging. As a result, the number of required datasets often seems insufficient to fully capture variations in both shape and appearance, especially in case of 3D modeling [13]. Various works such as FEM vibrational modes [4] have been proposed to artificially produce additional modes of variation, but it seems that this approach is mainly restricted to cope with intra-subject variability.

In this paper, we propose to study various shape and appearance priors in the context of lower limb bones segmentation using MR images and few training datasets. Two training sets are tested: one with only three datasets featuring a rather homogenous intensity distribution and the other with three more datasets featuring MRI artifacts. For shape modeling, both PCA and shape memory strategies are tested. PCA-based methods are known to need several datasets to be meaningful, while shape memory method requires in practice only one dataset. For appearance modeling, methods based on intensity profiles are tested, namely mean intensity profiles, multivariate Gaussian distributions of profiles and multimodal profiles from EM clustering. Our objective is to find the most efficient strategy, *i.e.* the strategy that is robust against the low number of datasets and their intensity inhomogeneities. This strategy would have the advantage to be more easily integrated in a clinical environment where the need of quick results is vital, regardless of the number of datasets.

2 MRI data

2.1 Acquisitions

For this study, six acquisitions were performed on six different subjects (4 females and 2 males, aged between 25 and 35). Protocols used for each acquisition are detailed in Table 1. The acquisitions took place in two different locations. Three subjects were scanned at the St Mary's Hospital, London, UK on a GE Medical Systems 1.5T MRI device (subjects #4, #5 and #6) and the three other subjects at the University Hospital of Geneva, Switzerland on a Philips Medical Systems 1.5T MRI device (subjects #1, #2 and #3). Both acquisitions feature 256x256 matrices but with a different slice thickness (*i.e.* 2 mm at London and 5 mm at Geneva). An institutional medical-ethical committee approved this study and subjects gave their written consent.

Table 1 MRI protocols used to scan the 6 subjects.

Subject	TR/TE(ms)	FOV(cm)/Matrix	Resolution(mm)
#1	4.15/1.69	35/256x256	1.37x1.37x5
#2	4.15/1.69	35/256x256	1.37x1.37x5
#3	4.15/1.69	35/256x256	1.37x1.37x5
#4	5.06/2.23	43/256x256	0.84x0.84x2
#5	4.34/1.56	40/256x256	0.78x0.78x2
#6	5.09/2.22	43/256x256	0.84x0.84x2

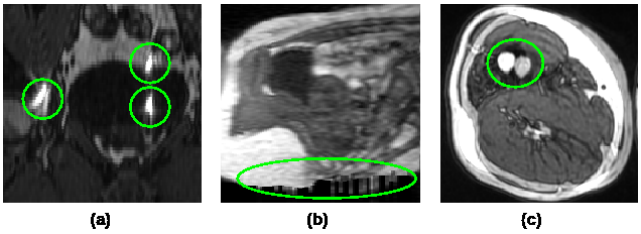

Fig. 1 Some MRI artifacts: (a) surface coil artifact on three distinctive spots, (b) slice-to-slice interference artifact at hip level and (c) subject motion artifact at the overlap between two consecutive slabs where the femur bone is clearly shifted.

Some MR images present strong artifacts like subject motion, surface coil, slice-to-slice interference and bias field (see Fig. 1). Subject motion is an artifact created by the displacement of a structure, which arises when the subject slightly moves between two consecutive acquisitions. Surface coil is characterized by a very strong signal due to the close proximity of the subject with the surface coil. Slice-to-slice interference artifact is due to the cross-excitation of adjacent slices with contrast loss in reconstructed images. Finally, bias field is a very common artifact in MR images, which may be induced by a number of factors such as poor radio frequency coil uniformity, static field inhomogeneity and radio frequency penetration.

2.2 Fusion of MRI slabs

An MRI acquisition of the whole lower limb cannot be performed in one scan due to a limited Field of View (FoV) of the machine. Consecutive scans, known as *slabs*, are necessary to cover its entire length. The registration of those slabs is then required to generate a complete MR image of the lower limbs. The registration is computed thanks to a sufficient overlap, which is used to compute the transformation matrix between slabs. In our case, we use a rigid registration based on the manual placement of landmarks.

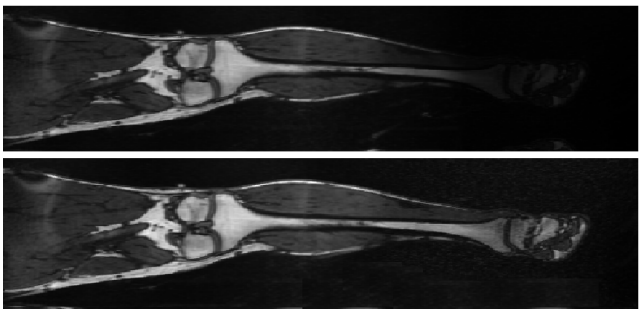

Fig. 2 MRI saggital view of the lower limbs before (top) and after (bottom) intensity correction. Note how the intensity has been corrected at foot level.

However, the registration is not enough to create a satisfactory MR image of the lower limbs. Indeed, intensity distribution may vary between registered slabs (see top of Fig. 2), which means the intensity histogram for the same structure may slightly differ between slabs. This is because the intensity range of the MR image may change between slabs (*i.e.* the minimum and maximum of intensity value is different). Also, the presence of strong artifacts at the image boundaries is likely to affect the intensity distribution (in addition to the common bias field). These artifacts are unpredictable and may differ in their number and importance between slabs.

Our solution consists in putting artifact intensity values into the background, so that when normalizing intensity between slabs, artifacts do not bias the correction. The normalization is performed using an intensity scaling factor. To this end, we calculate for each slab the histogram of the main structures (*i.e.* bones, muscles and fat) and compute the scaling factor so that histograms are similar. Finally, a bias correction is applied on the generic MR image of the lower limbs to correct intensity inhomogeneities [23], *i.e.* to correct the bias field and to smooth the intensity scaling between slabs (see bottom of Fig. 2).

3 Creation of appearance and shape priors

In this paper, we propose to exploit priors built from both the appearance and shape of structures in model-based image segmentation. A set of P training shapes $\{\mathbf{S_1}, \ldots, \mathbf{S_P}\}$ with corresponding images $\{I_1, \ldots, I_P\}$ is necessary to model the priors.

Each shape $\mathbf{S}$ is modeled as a 2-simplex mesh [8] defined by N points $\mathbf{x}_i$ with normals n_i:

$$\mathbf{S} = \{(\mathbf{x}_1, n_1), \ldots, (\mathbf{x}_N, n_N)\}$$

The shapes are produced by a supervised segmentation approach in which the point correspondence is established (*i.e.* landmarks on all P training shapes are located at corresponding positions). This point correspondence, which is the first necessary step when building shape models with PDM, is ensured through a registration between training shapes. Then, the P training shapes are aligned in a common coordinate frame. The most popular method to solve this problem is the General Procrustes Analysis (GPA) [11, 12], which aligns the set of P training shapes to their unknown mean by minimizing the mean squared distance between two shapes in an iterative procedure. After alignment, dimensionality of the training set is reduced to find a small set of modes that best describes the observed variation. This is accomplished using PCA [16]. For more information on the issues of building training sets, we refer the interested readers to the recent review of Heimann *et al* [13].

3.1 Shape priors

For the shape prior construction, we propose to use both PCA-based priors (PCA) and shape memory (SMEM).

3.1.1 PCA

After the alignement of the P training shapes into a common coordinate frame with GPA, a Statistical Shape Model (SSM) is built [7].

An arbitrary shape $\mathbf{S}$ is approximated from the computed statistics by:

$$\mathbf{S} \approx T(\overline{\mathbf{S}} + \Phi\, \mathbf{b})$$

where vector $\overline{\mathbf{S}}$ is the mean shape, Φ is a matrix of M ($M \leq P$) principal modes (with respective variances λ_i), $\mathbf{b}$ is a vector of shape parameters and T denotes the alignment transform.

To estimate the unknown parameters $\mathbf{b}$ and T, an iterative procedure is used [6]. To ensure the SSM specificity, two kinds of constraints are considered: *hard* or *soft*.

Hard constraint is defined as:

$$-3\sqrt{\lambda_m} \leq b_m \leq 3\sqrt{\lambda_m}, \forall m \in [1, M]$$

Soft constraint scales b_m to lie inside a hyperellipsoid:

$$\sum b_m^2/\lambda_m \leq C, \forall m \in [1, M]$$

where C is computed from the χ^2_M distribution [6].

When a shape is replaced by its closest shape counterpart, we qualify this replacement as *PCA regularization*.

3.1.2 SMEM

A single shape can also express a basic prior by using some of its geometrical properties. In case of a 2-simplex mesh, a local description of each vertex with respect to its three neighbors is computed [8]. Only three independent simplex parameters are necessary, and those are similarity transform invariant. By using these prerecorded parameters, shape can be locally recovered. Although this prior is based on a single *representative* shape, the similarity invariance property confers more flexibility than encoding the shape as 3D points. This simple prior based on the simplex parameters is often denoted as *shape memory* and is similar to the notion of strain energy in mechanics.

3.2 Appearance priors

To take appearance into account, intensity profiles $\mathbf{p_i}$ are built by sampling the image intensity at each point $\mathbf{x}_i$ along the normal direction [7, 8]. From these profiles, various approaches to build a prior have been presented in the literature. In this paper, we consider mean intensity profiles (PROF), multivariate Gaussian distributions of intensity profiles (MGD) and multimodal profiles (MPAM), which are built from an EM clustering of intensity profiles (see Fig. 3).

3.2.1 PROF

Mean intensity profiles constitute the simplest appearance prior. At each corresponding point through all datasets, a mean intensity profile is computed as:

$$\boldsymbol{\mu_i} = \sum_{j=1}^{P} \mathbf{p_i}/P$$

Though faster to compute due to its simplicity, this prior does not make any assumption about the variance. As a result, mean intensity profiles are rather poor priors. However, they have been successfully exploited in previous works [10, 19] when combined with robust similarity measures such as the Normalized Cross Correlation (NCC) [15].

3.2.2 MGD

Cootes *et al.* [7] proposed to model the prior of the (normalized) intensity profile at each corresponding point by a multivariate Gaussian of mean $\boldsymbol{\mu_i}$ and covariance matrix Σ_i. This model has been successfully exploited in various works [6, 13]. As $\boldsymbol{\mu_i}$ and Σ_i are provided by the prior, the Mahalanobis distance is usually used as similarity measure.

However, such a distance needs the computation of the inverse of Σ_i. This computation is known to be problematic, since the inversion of the covariance matrix may lead to singular matrices. This is especially the case with intensity profiles that feature a high dimensional space (*i.e.* profiles are large to efficiently describe appearance), but without enough profiles to represent those dimensions (so called *curse of dimensionality*). In our case, this effect is even more important as we study statistical models performance when using few training datasets.

In the literature, several techniques have been proposed to regularize the covariance matrix, and thus to avoid singularities [24]. Cootes proposed an alternative approach [6] consisting in first performing a PCA on Σ_i, and then computing the Mahalanobis distance directly from the PCA statistics using the principal modes and associated eigenvalues. With such approach, no matrix inversion needs to be performed. In addition, a dimensionality reduction can be applied by selecting only the first modes, which leads to an approximated Mahalanobis distance.

3.2.3 MPAM

The clustering of intensity profiles is formulated in the context of a probability density estimation using Gaussian Mixture Models (GMM) [3]. Intensity profile classes are estimated using the Expectation-Maximization algorithm (EM) for each mesh but not for each point (*i.e.* without the need for any registration). EM is initialized with the Fuzzy C-Means algorithm (FCM), which is itself initialized with random cluster centers.

A large profile length (to efficiently describe appearance) combined with a coarse sampling of meshes (to speed up computation) is likely to lead to singular matrices when inverting EM covariance matrices during the E-step. To solve this problem, three distinct methods to regularize the covariance matrix are proposed: *Spectral*, *Diagonal* and *Constant Regularization*. To determine the number of classes that best represents the data, a novel non parametric model order selection criterion called *Overlap Separation Index* (OSI) inspired by cluster validity indices [17] is used.

Spatially, the EM classification does not take into account the neighborhood information of intensity profiles. This leads to a non spatially smooth distribution of intensity profiles into the different clusters, which may impair the fusion of appearance regions. To take the connectivity between profiles into account, the Neighborhood EM algorithm (NEM) [1] is used. Fig. 3 outlines an example of those different steps used for intensity profile clustering.

After the classification, each mesh $\mathcal{M}_p$ is associated with $\mathcal{K}_p$ clusters. MPAM are created by the fusion of these $\mathcal{K}_p$ clusters that may vary among meshes. The $\mathcal{K}_p$ clusters are compared between meshes and possibly merged. In order to merge similar clusters, the Jaccard index is used as a sim-

ilarity measure. A threshold $\mathcal{J}$ between 0 and 1 is used to decide whether two clusters are equivalent.

The last step consists in providing a geometric embedding for the new clusters. For a set of meshes without point correspondence, each mesh $\mathcal{M}_p$ needs to be registered to a reference mesh $\mathcal{M}^\star$, on which each posterior probability is resampled using a closest point approach [3]. However, since we use a set of P meshes $\mathcal{M}_p$ with point correspondence, this registration is not necessary.

At the end, several clusters, or *intensity profile priors*, may be assigned at each vertex of reference mesh $\mathcal{M}^\star$, which leads to a multimodal approach.

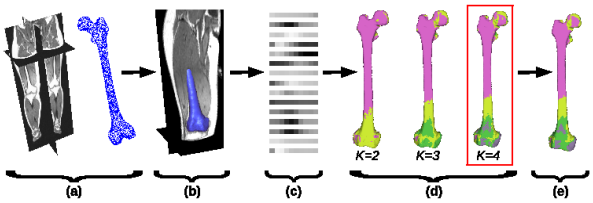

Fig. 3 Construction of the MPAM appearance prior of a femur bone. From an image and a mesh (a), a segmentation is performed (b). Intensity profiles are sampled (c), then classified (d) into classes (here, K={2,3,4}). Using an appropriate criterion, best classification is chosen (here, for K=4). Finally, the classification is spatially smoothed (e).

4 Segmentation based on priors

The proposed segmentation is based on dynamic deformable models [10, 19]. In such segmentation, a deformable template evolves until reaching an equilibrium. Internal forces regulate its evolution while external forces drive it towards anatomical boundaries. This section discusses the effect of abovementioned priors into this deformable model framework.

4.1 Evolution

Dynamic deformable models behave like a particle system, in which each particle corresponds to a lumped-mass vertex subject to internal and external forces. The dynamics of the system follow the Newtonian laws of motion, which express particle state (*i.e.* position and velocity) with respect to forces. The resulting time-discretized differential equation system is solved by an implicit Euler scheme.

A multiresolution scheme is used to produce various levels of detail (LOD) of the shapes [10] (see Fig. 5). The LOD are then exploited in a coarse-to-fine fashion, improving the robustness and accuracy of the segmentation evolution. In case of simultaneous segmentation of more than one structure, efficient collision response and detection are applied to prevent interpenetrations [10]. Alternatively, a post-processing correction method can be used [20]. Forces are expressed based on the image information, the current model configuration and the pre-computed priors.

Forces at point $\mathbf{x}_i$ are expressed as the force of a Hookean spring:

$$f_i = \alpha * (\tilde{\mathbf{x}}_i - \mathbf{x}_i)$$

where $\tilde{\mathbf{x}}_i$ denotes the *target* point and α is a weighting coefficient specific to each type of force. We will see in the following how the target point is computed given the different forces. This procedure is hereupon referred to as *source-to-target* approach.

4.2 Internal forces based on shape priors

4.2.1 PCA

As depicted in section 3.1.1, the shape priors are expressed by a SSM built on a PCA. At each iteration, a closest shape $\hat{\mathbf{S}}$ is found by projecting the current deformable shape $\mathbf{S}$ into the PCA space. An iterative procedure computes the adequate transformation T and appropriate constrained shape parameters $\mathbf{b} = b_1, \ldots, b_M$ [6].

Hard or soft constraints are applied to discard illegal configurations. Then, $\hat{\mathbf{S}} = \{\hat{x_1}, \ldots, \hat{x_N}\}$ eventually becomes the target shape and the source-to-target approach is applied:

$$f_i^{\text{pca}} = \alpha^{\text{pca}}(\hat{\mathbf{x}}_i - \mathbf{x}_i)$$

4.2.2 SMEM

In case of a single reference shape used as a prior, both pre-computed and current simplex parameters are used to derive new target point positions [8, 10]. A force f_i^{smem} is then produced at each point $\mathbf{x}_i$.

4.3 External forces based on appearance priors

At each iteration of the evolution, a number W of intensity profiles $\{\mathbf{p_i}^1, \ldots, \mathbf{p_i}^W\}$ are sampled along the normal n_i at point $\mathbf{x}_i$. Among them, a target profile $\tilde{\mathbf{p}}_\mathbf{i}$ is chosen, whose corresponding position is $\tilde{\mathbf{x}}_i$. Usually, this target profile is chosen so that it maximizes a similarity criterion or minimizes a distance.

4.3.1 PROF

In case of mean intensity profiles, a target profile $\tilde{\mathbf{p_i}}$ is selected if it maximizes the Normalized Cross Correlation NCC with the mean intensity profile $\boldsymbol{\mu_i}$:

$$\tilde{\mathbf{p_i}} = \text{argmax } NCC(\mathbf{p_i}^j, \boldsymbol{\mu_i}) \quad \text{where } j \in [1, W]$$

4.3.2 MGD

When using a multivariate Gaussian distribution of intensity profiles, the information from the covariance matrix Σ_i is also taken into account. In this case, a target profile $\tilde{\mathbf{p_i}}$ is selected if it minimizes the Mahalanobis distance d_M derived from the computed distribution [6]:

$$\tilde{\mathbf{p_i}} = \text{argmin } d_M(\mathbf{p_i}^j) = (\mathbf{p_i} - \boldsymbol{\mu_i})^T \Sigma_i^{-1} (\mathbf{p_i} - \boldsymbol{\mu_i})$$

where $j \in [1, W]$.

As previously said in section 3.2.2, a regularization of the covariance matrix Σ_i is necessary to avoid any singularities [18]. However, the Mahalanobis distance d_M may be computed without inverting the covariance matrix, *i.e.* by using an alternative approach proposed by Cootes [6].

Given a PCA performed on P profiles $\mathbf{p_i}$ at vertex $\mathbf{x}_i$, expressed by the principal matrix Φ_i, the m eigenvalues λ_{il} and the mean intensity profile $\boldsymbol{\mu_i}$, Mahalanobis distance d_M is defined as:

$$d_M(\mathbf{p_i}) = \sum_{l=1}^{m} \frac{b_{il}^2}{\lambda_{il}}$$

where $b_i = (b_{i1}, \ldots, b_{im})^T$ is the model parameter vector of the best fit $\hat{\mathbf{p}}_\mathbf{i}$ of $\mathbf{p_i}$ given the PCA model:

$$\hat{\mathbf{p}}_\mathbf{i} = \boldsymbol{\mu_i} + \Phi_i b_i$$

4.3.3 MPAM

In case of multimodal profiles, the comparison with multiple intensity profile priors is considered. Indeed, since the prior is multimodal, more than one intensity profile prior may be assigned to $\mathbf{x}_i$. In practice, only intensity profile priors whose posterior probability is higher than a threshold (*e.g.* 10^{-3}) are considered. This can be done because EM classification leads to sharp posterior probabilities whose values are either very high or very low (*i.e.* quite below 10^{-3}).

Similarly to MGD, the Mahalanobis distance is used as similarity measure. But this time, a profile $\mathbf{p_i}$ is compared with the K_i intensity profile priors associated to $\mathbf{x}_i$ (*i.e.* remember that each point $\mathbf{x}_i$ is associated with a series of clusters, whose number may not be the same for every point). Each point $\mathbf{x}_i$ is associated with K_i clusters of center $\boldsymbol{\mu}_\mathbf{i}^\mathbf{k}$ and covariance matrix Σ_i^k.

The target profile $\tilde{\mathbf{p_i}}$ of a given point $\mathbf{x}_i$ is chosen as one of the W profiles sampled along the normal that has the smallest Mahalanobis distance d_M with one of its K_i associated clusters:

$$\tilde{\mathbf{p_i}} = \text{argmin } d_M(\mathbf{p_i}^{j,k})$$

where $j \in [1, W]$ and $k \in [1, K_i]$.

Unlike MGD, only the diagonal terms $\{\sigma_{i,1}^k, \ldots, \sigma_{i,L}^k\}$ of the $L \times L$ covariance matrix Σ_i^k are taken into account:

$$d_M(\mathbf{p_i}^k) = (\mathbf{p_i} - \boldsymbol{\mu_i})^T \text{diag}(1/\sigma_{i,1}^k, \ldots, 1/\sigma_{i,L}^k)(\mathbf{p_i} - \boldsymbol{\mu_i})$$

where $\text{diag}(1/\sigma_{i,1}^k, \ldots, 1/\sigma_{i,L}^k)$ depicts a diagonal matrix.

Using only the diagonal terms has the advantage to speed up computation, which may be considerable as the comparison is now multimodal. This makes also sense since the covariance matrix regularization performed during MPAM creation strongly reduces the influence of non-diagonal elements.

5 Experimental setup

Two training sets, D3 and D6 with three and six lower limb MRI datasets respectively (see Fig. 4 for more details), are used with Leave-One-Out (LOO) cross validation. A multiresolution scheme is used to produce four levels of detail for each structure. For femur bone, the four resolutions consist in $N = 514$, $N = 2056$, $N = 8224$ and $N = 32896$ vertices respectively. For hip bone, they consist in $N = 814$, $N = 3256$, $N = 13024$ and $N = 52096$ vertices respectively (see Fig. 5).

Three appearance models are compared: mean intensity profiles (PROF) with Normalized Cross Correlation measure, multivariate Gaussian distributions of intensity profiles (MGD) and multimodal profiles (MPAM), both with Mahalanobis distance. With MGD, PCA takes 95% of the total intensity into account to compute the approximated Mahalanobis distance. To regularize the covariance matrix during EM iterations, MPAM is created using *Constant Regularization* method coupled with parameter h = 1.0 [3]. Due to the limited number of training datasets, we prefer not to merge any mode (*i.e.* Jaccard index $\mathcal{J}$ = 1.0).

Regarding intensity profiles, thirty-one intensities are sampled every 0.5 mm from 12.5 mm inside to 2.5 mm outside mesh surface at each vertex for PROF and MGD, since those values were successfully used for the segmentation of bones

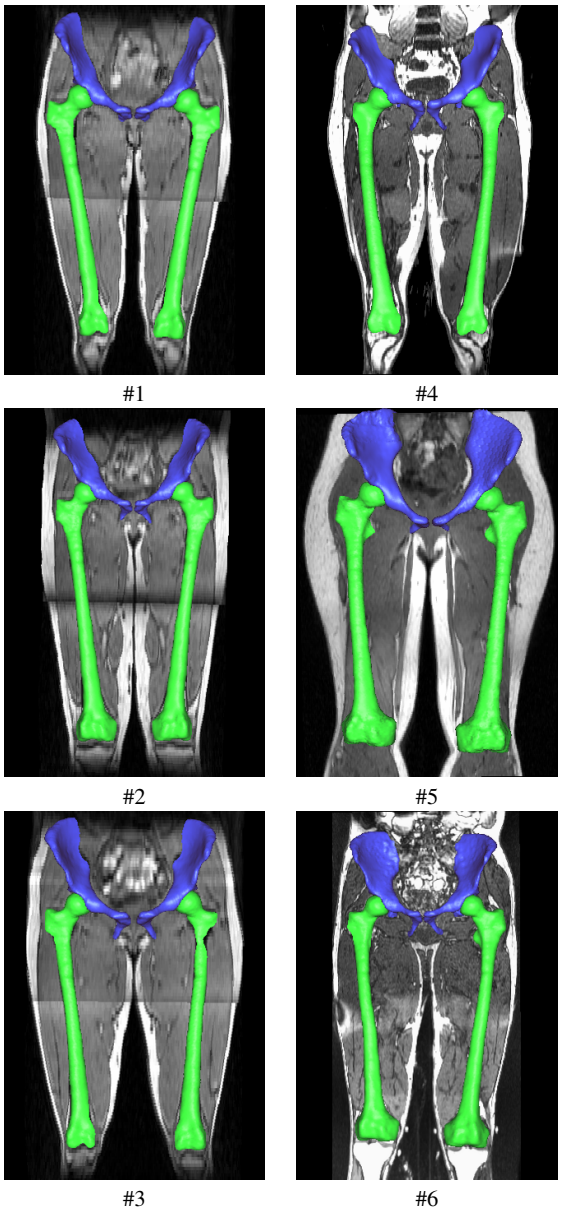

Fig. 4 The six datasets that are used for the experiments with their image and corresponding reference meshes (*i.e.* hip bones in blue and femur bones in green). The three datasets on the left constitute the D3 training set while all datasets together constitute the D6 training set. Note how rather linear is the intensity distribution between datasets of D3. Conversely, the three datasets on the right feature a quite heterogeneous intensity distribution. Among them, dataset #4 and dataset #6 are strongly subject to MRI artifacts.

in MR images [10, 19]. For MPAM, appropriate values need to be defined since the appearance prior construction method is different. Experiments showed that eleven intensities sampled every mm from 5 mm inside to 5 mm outside mesh surface give reasonable results in terms of clustering and segmentation. Those values are thus used in the framework of this work.

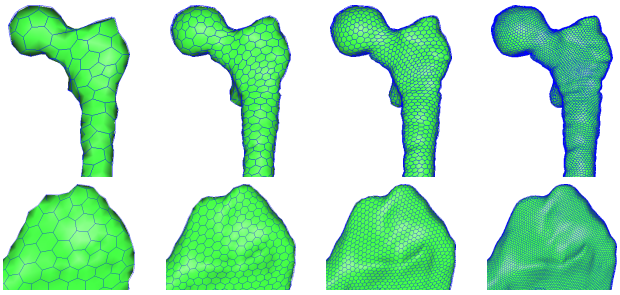

Fig. 5 Detail of the four increasing resolutions (from left to right) for both femur (top) and hip (bottom) bones. For femur bone, the four resolutions are $N = 514$, $N = 2056$, $N = 8224$ and $N = 32896$ vertices respectively. For hip bone, the four resolutions are $N = 814$, $N = 3256$, $N = 13024$ and $N = 52096$ vertices respectively.

Table 2 Mean DICE measure (*i.e.* on all structures and on all segmented datasets) when combining appearance and shape priors on D3 training set.

	MGD	PROF	MPAM
SMEM	88.48	91.98	87.39
PCA	78.82	89.51	89.40

To have a fair comparison between methods, same initialization and internal forces (*i.e.* PCA-based prior or shape memory) are used for every appearance model. The initialization is based on the manual placement of landmarks coupled with a shape interpolation approach [10]. Segmentation accuracy is assessed with DICE measure (DSC) [9] based on reference manual segmentations. In the results, femur and hip bones are both considered as one structure, though they are both represented by two instances (*i.e.* left and right). To simplify the statistical analysis, the DICE measure on each structure is actually a mean on its both instances. Also, only datasets #1, #2 and #3 are segmented, since they feature a more homogeneous intensity distribution (see Fig. 4).

6 Results

First results consist in averaging the DICE measure on all structures and on all segmented datasets. Using D3 (see Table 2), PROF gives the best results, with both SMEM (DSC = 91.98) and PCA (DSC = 89.51), followed closely by MPAM with PCA (DSC = 89.40). Results are quite similar using D6 (see Table 3), PROF performs the best with both SMEM (DSC = 90.91) and PCA (DSC = 89.14), followed by MPAM with PCA (DSC = 88.08). MGD coupled with PCA gives by far the worst results, when using both D3 and D6. Regardless of the appearance prior used, SMEM is more accurate than PCA except when the latter is coupled with MPAM. When comparing global results on D3 and D6, we notice that D3 always gives better results.

In a second step, we study in more details the results on the two separate structures (see some delineations on Fig. 6

Table 3 Mean DICE measure (*i.e.* on all structures and on all segmented datasets) when combining appearance and shape priors on D6 training set.

	MGD	PROF	MPAM
SMEM	87.64	90.91	87.19
PCA	70.37	89.14	88.08

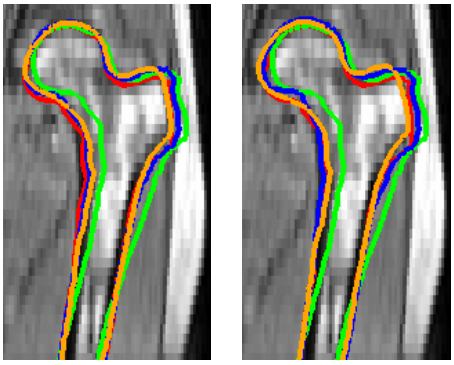

Fig. 6 Segmentation of left femur bone from dataset #1 using D3 training set. Results of PROF (left) and MPAM (right) appearance priors. Reference mesh is depicted in red, initialization in green, SMEM shape prior in blue and PCA shape prior in orange.

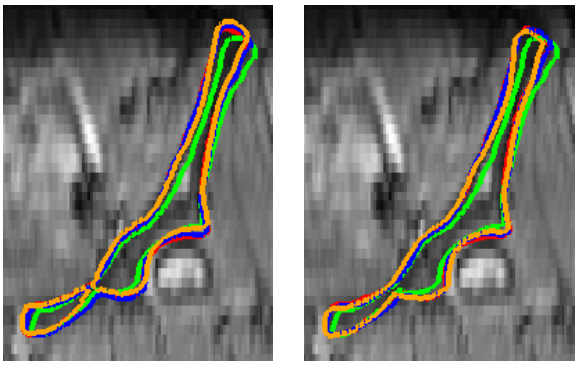

Fig. 7 Segmentation of left hip bone from dataset #1 using D3 training set. Results of MGD (left) and PROF (right) appearance priors. Reference mesh is depicted in red, initialization in green, SMEM shape prior in blue and PCA shape prior in orange.

and Fig. 7 for femur and hip bones respectively). Using D3 (see Table 4), PROF gives the best results with SMEM for femur bone (DSC = 92.51). This combination performs also the best for hip bone (DSC = 91.45). Using D6 (see Table 5), MPAM coupled with PCA gives the most accurate segmentation for femur bone (DSC = 92.02). However, for hip bone, PROF with SMEM gives once again the best results (DSC = 90.09). A clear tendency shows that results for femur bone are clearly better than for hip bone, regardless of the training set, shape and appearance prior used.

When comparing the influence of shape priors on femur bone segmentation with D6, we notice very slight differences between appearance priors, except for MPAM. SMEM performs the best for hip bone segmentation though (with both D3 and D6), but once again except for MPAM. When comparing the influence of the training set, D3 gives better results on hip bone, regardless of the shape and appearance prior used. Conversely, D6 gives better results on femur bone when PCA is used with all appearance priors.

Table 4 Mean $\pm$ standard deviation of DICE measure for each structure and on all segmented datasets when combining appearance and shape priors on D3 training set.

Femur	MGD	PROF	MPAM
SMEM	92.28 ± 0.63	92.51 ± 1.36	90.51 ± 0.51
PCA	90.57 ± 1.36	90.52 ± 0.91	90.71 ± 0.90
Hip	**MGD**	**PROF**	**MPAM**
SMEM	84.67 ± 9.84	91.45 ± 2.46	84.27 ± 6.58
PCA	67.07 ± 30.6	88.50 ± 2.07	88.10 ± 3.22

Table 5 Mean $\pm$ standard deviation of DICE measure for each structure and on all segmented datasets when combining appearance and shape priors on D6 training set.

Femur	MGD	PROF	MPAM
SMEM	91.92 ± 0.89	91.74 ± 1.72	90.46 ± 0.37
PCA	91.97 ± 1.42	91.62 ± 1.71	92.02 ± 0.96
Hip	**MGD**	**PROF**	**MPAM**
SMEM	83.36 ± 5.24	90.09 ± 2.22	83.92 ± 5.65
PCA	48.77 ± 33.1	86.66 ± 2.29	84.13 ± 7.20

7 Discussion

Generating statistics from a segmentation framework using a small amount of training datasets is a quite challenging work and a difficult task. Indeed, statistics are usually relevant when numerous data are at disposal. However, results on both training sets D3 and D6 show some trends that we discuss in this section.

In general, using only three datasets with D3 gives better results than six datasets with D6. This may appear as counter-intuitive: more datasets should lead to better results. However, the three additional datasets from D6 are mostly corrupted by noise and artifacts. They also feature a quite heterogeneous intensity distribution. In this case, adding more information, which is corrupted, does not improve the appearance prior but rather weakens it. As a result, external forces have more difficulties to find the right boundaries.

In general, using shape memory with SMEM gives better results than PCA-based priors. This is because SMEM is a local approach. Given a shape that does not exhibit large inter-subject variation (*i.e.* like bones compared to soft organs) and that is sufficiently initialized, SMEM is a quite robust approach that can deal with few training datasets. However, PCA is known to give better results when using more datasets. In fact, a PCA-based evolution tends to be less sensitive to initialization when a sufficient number of datasets is available to provide a richer information about the shape [21]. This is explained by the fact that PCA af-

fects all the shape vertices in a global manner compared to the shape memory force, which confers a better robustness against local artifacts.

As depicted in both Table 4 and Table 5, MGD coupled with PCA gives very bad results when segmenting hip bone (DSC = 67.07 with D3 and 48.77 with D6). This would suggest that this combination is the worst: PCA as a shape prior (in case of few training datasets) and MGD as an appearance prior. External forces generated by MGD are thus less efficient than those generated by PROF and MPAM. This would explain the huge difference of DSC between D3 and D6, knowing that D6 contains datasets with images corrupted by noise and artifacts. As a result, those external forces are more sensitive to internal forces. But since internal forces based on PCA are also weak due to the few number of datasets, the segmentation is doomed to give bad results.

Regarding MPAM, the best results are for femur bone segmentation when combined with PCA and when using D6. Though in theory MPAM should need less datasets than MGD (*i.e.* a PCA-based appearance method), MPAM seems to feature a certain sensitivity to noise. Indeed, results for femur bone are clearly better than for hip bone (when using both D3 and D6). This is because intensity distribution is more heterogeneous at hip level, as hip bones are located close to image top border. This would suggest that the similarity measure (*i.e.* the Mahalanobis distance) should be optimized to cope with non linear intensity. Moreover, the intensity profile computation is different for MPAM than for PROF and MGD. Additional experiments should be performed to determine an optimal intensity profile length for the segmentation of bones in MR images.

In general, PROF gives the best results (*i.e.* compared to MGD and MPAM). We believe there are two explanations for that. First, reference meshes are produced by a supervised approach that uses a semi-automatic segmentation controlled by manually defined constraints [21]. This semi-automatic segmentation exploits the same deformable model evolution coupled with the NCC similarity measure reported in this paper. This creates a bias, which is likely to give an advantage to PROF-based appearance prior (*i.e.* PROF also uses NCC during the segmentation). Second, NCC similarity measure is known to be more robust to intensity change (*i.e.* NCC is affine invariant). Using NCC would thus help PROF to be more efficient in presence of intensity change (*i.e.* especially with the three additional datasets corrupted by noise and artifacts that are used with D6).

8 Conclusion

When using a small amount of training datasets for a segmentation, results tend to show that local and simple methods perform the best. As a shape prior, shape memory (SMEM) gives very good results. As an appearance prior, mean intensity profiles (PROF) gives the best results. We believe these good results are partly due to the Normalized Cross Correlation (NCC) similarity measure, which is more robust to intensity change in MR images.

When increasing the number of training datasets, results tend to show that statistical methods feature the best results: PCA-based shape prior (PCA) and multimodal profiles (MPAM) as an appearance prior. Both methods capture more and more information while the number of training datasets is increased and we believe that better segmentation results would be obtained when increasing this number of datasets. However, this hypothesis only holds when the training data is of enough quality to produce meaningful and efficient priors.

Acknowledgment This work is supported by the 3D Anatomical Human project (MRTN-CT-2006-035763) funded by the European Union. The authors would like to thank the six subjects that took part in the MRI acquisitions and both St Mary's Hospital and University Hospital of Geneva. Special thanks to Mitchelle Chen and Prof. Andrew Todd-Pokropek from University College London for their help and collaboration during the MRI acquisitions in London.

References

1. Ambroise C, Dang M, Govaert G (1997) Clustering of spatial data by the em algorithm. Quantitative Geology and Geostatistics 9:493–504
2. Behiels G, Maes F, Vandermeulen D, Suetens P (2002) Evaluation of image features and search strategies for segmentation of bone structures in radiographs using active shape models. Medical Image Analysis 6(1):47–62
3. Chung F, Delingette H (2009) Multimodal prior appearance models based on regional clustering of intensity profiles. In: MICCAI 2009 - Proceedings of the 12th International Conference on Medical Image Computing and Computer Assisted Intervention, Lecture Notes in Computer Science, vol 5762, pp 1051–1058
4. Cootes T, Taylor C (1996) Data driven refinement of active shape model search. In: BMVC 1996 - Proceedings of the 7th British Machine Vision Conference
5. Cootes T, Taylor C (2004) Statistical models of appearance for computer vision
6. Cootes TF, Hill A, Taylor CJ, Haslam J (1993) The use of active shape models for locating structures in medical images. In: IPMI'93 - Proceedings of the 13th International Conference on Information Processing in Medical Imaging, pp 33–47

7. Cootes TF, Edwards G, Taylor CJ (2001) Active appearance models. IEEE Pattern Analysis and Machine Intelligence 23(6):681,685
8. Delingette H (1999) General object reconstruction based on simplex meshes. International Journal of Computer Vision 32(2):111–146
9. Dice L (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302
10. Gilles B, Magnenat-Thalmann N (2010) Musculoskeletal mri segmentation using multi-resolution simplex meshes with medial representations. Medical Image Analysis 14(3):291–302
11. Goodall C (1991) Procrustes methods in the statistical analysis of shape. Journal of the Royal Statistical Society Series B (Methodological) 53(2):285–339
12. Gower J (1975) Generalized procrustes analysis. Psychometrika 40:33–51
13. Heimann T, Meinzer HP (2009) Statistical shape models for 3d medical image segmentation: A review. Medical Image Analysis 13:543–563
14. Heimann T, Munzing S, Meinzer H, Wolf I (2007) A shape-guided deformable model with evolutionary algorithm initialization for 3d soft tissue segmentation. In: Karssemeijer N, Lelieveldt B (eds) IPMI 2007 - Proceedings of the 20th International Conference on Information Processing in Medical Imaging, vol 4584, pp 1–12
15. Holden M, Hill D, Denton E, Jarosz J, Cox T, Hawkes D (1999) Voxel similarity measures for 3d serial mr brain image registration. In: IPMI'99 - Proceedings of the 16th International Conference on Information Processing in Medical Imaging, Springer, vol LNCS 1613, pp 472–477
16. Jolliffe IT (2002) Principal Component Analysis, 2nd edn. Springer
17. Kim DW, Lee KH, Lee D (2004) On cluster validity index for estimation of the optimal number of fuzzy clusters. Pattern Recognition 37:2009–2025
18. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4(1)
19. Schmid J, Magnenat-Thalmann N (2008) Mri bone segmentation using deformable models and shape priors. In: MICCAI 2008 - Proceedings of the 11th International Conference on Medical Image Computing and Computer Assisted Intervention, Lecture Notes in Computer Science, vol 5241, pp 119–126
20. Schmid J, Sandholm A, Chung F, Thalmann D, Delingette H, Magnenat-Thalmann N (2009) Musculoskeletal simulation model generation from MRI datasets and motion capture data, Springer-Verlag, pp 3–19
21. Schmid J, Kim J, Magnenat-Thalmann N (2010) Extreme leg motion analysis of professional ballet dancers via mri segmentation of multiple leg postures. International Journal for Computer Assisted Radiology and Surgery In Press available online
22. Seim H, Kainmueller D, Heller M, Lamecker H, Zachow S, Hege HC (2008) Automatic segmentation of the pelvic bones from ct data based on a statistical shape model. In: Botha C, Kindlmann G, Niessen W, Preim B (eds) Eurographics Workshop on Visual Computing for Biomedicine, Eurographics Association, pp 93–100
23. Styner M, Brechbühler C, Székely G, Gerig G (2000) Parametric estimate of intensity inhomogeneities applied to mri. IEEE Transactions on Medical Imaging 19:153–165
24. Tadjudin S, Landgrebe D (1999) Covariance estimation with limited training samples. IEEE Transactions on Geoscience and Remote Sensing 37(4):2113–2118